doi:10.1016/j.jmb.2004.12.039 J. Mol. Biol. (2005) 346, 1083–1093

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

The Structure of the Neurotoxin-associated Protein HA33/A from *Clostridium botulinum* Suggests a Reoccurring β-Trefoil Fold in the Progenitor Toxin Complex

Joseph W. Arndt[1], Jenny Gu[1], Lukasz Jaroszewski[2] Robert Schwarzenbacher[2], Michael A. Hanson[1], Frank J. Lebeda[3] and Raymond C. Stevens[1]*

[1]*Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037 USA*

[2]*The Joint Center for Structural Genomics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

[3]*Department of Cell Biology and Biochemistry, Toxinology Division, U.S. Army Medical Research Institute of Infectious Diseases Frederick, MD 21702 USA*

**Corresponding author*

The hemagglutinating protein HA33 from *Clostridium botulinum* is associated with the large botulinum neurotoxin secreted complexes and is critical in toxin protection, internalization, and possibly activation. We report the crystal structure of serotype A HA33 (HA33/A) at 1.5 Å resolution that contains a unique domain organization and a carbohydrate recognition site. In addition, sequence alignments of the other toxin complex components, including the neurotoxin BoNT/A, hemagglutinating protein HA17/A, and non-toxic non-hemagglutinating protein NTNHA/A, suggests that most of the toxin complex consists of a reoccurring β-trefoil fold.

Keywords: neurotoxin; hemagglutinin; β-trefoil; progenitor toxin; sugar-binding

Introduction

The *Clostridium botulinum* neurotoxins (BoNTs) are among the most toxic bacterial toxins known. Exposure to toxin prevents the release of acetylcholine at neuromuscular junctions and synapses by cleaving one of the three neuronal proteins of the soluble *N*-ethylmaleimide-sensitive-factor attachment protein receptor (SNARE) complex required for synaptic vesicle membrane fusion resulting in flaccid muscle paralysis.[1–4] Ironically, these toxins are used clinically to treat neuromuscular disorders.

C. botulinum produces the 150 kDa BoNT concomitantly with a group of non-toxic neurotoxin-associated proteins (NAPs), forming complexes known as progenitor toxins. Previous studies have demonstrated that the NAPs protect BoNT from acid denaturation in the stomach and attack from a variety of proteolytic enzymes in the gastrointestinal tract.[5–8] Seven distinct BoNTs have been identified and are referred to as serotypes A–G, with serotype A being the most virulent to humans.[9] BoNT/A is secreted as a progenitor toxin complex in one of three sizes, 12 S (300 kDa), 16 S (500 kDa), or 19 S (900 kDa), depending on the type and number of NAPs associated with the complex.[3,10] The 12 S progenitor toxin complex consists of a single BoNT molecule and one non-toxic non-hemagglutinin (NTNHA) protein, but lacks the associated proteins responsible for hemagglutination activity. The 16 S progenitor toxin, in

Present address: J. Gu, San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

Abbreviations used: HA, hemagglutinin; NTNHA, non-toxic non-hemagglutinin; BoNT, botulinum neurotoxin; NAP, neurotoxin-associated protein; RMSD, root-mean-square deviation; TeNT, tetanus neurotoxin.

E-mail address of the corresponding author: stevens@scripps.edu

0022-2836/$ - see front matter

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1 JAN 2005	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The structure of the neurotoxin- associated protein HA33/A from Clostridium botulinum suggests a reoccurring beta-trefoil fold in the progenitor toxin complex, Journal of Molecular Biology 346:1083 - 1093	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Arndt, JW Gu, J Jaroszewski, L Schwarzenbacher, R Hanson, MA Lebeda, FJ Stevens, RC	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
The hemagglutinating protein HA33 from Clostridium botulinum is associated with the large botulinum neurotoxin secreted complexes and is critical in toxin protection, internalization, and possibly activation. We report the crystal structure of serotype A HA33 (HA33/A) at 1.5 A resolution that contains a unique domain organization and a carbohydrate recognition site. In addition, sequence alignments of the other toxin complex components, including the neurotoxin BoNT/A, hemagglutinating protein HA17/A, and non-toxic non-hemagglutinating protein NTNHA/A, suggests that most of the toxin complex consists of a reoccurring beta-trefoil fold.

15. SUBJECT TERMS
Clostridium botulinum, neurotoxin, BOT, hemagglutination, crystal structure, sequence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**11**	

addition to the components found in the 12 S complex, contains three hemagglutinin (HA) proteins, HA70 (also referred to as HA3a and HA3b after proteo-lysis), HA33, and HA17. The secreted and most toxic form 19 S complex is believed to be a dimer of two 16 S toxins linked by an additional HA33 protein.[11]

The HA positive progenitor toxins have been shown to bind to glycolipids and glycoproteins of the intestinal microvilli through interactions with oligosaccharides, facilitating internalization and transport in the bloodstream.[12–14] Interestingly, the oral ingestion of the progenitor complexes displays ~100 times more toxicity than the naked BoNT protein.[15] Though the molecular details leading to the higher efficacy of the progenitor toxin complexes remain unknown, it has been proposed that their improved effectiveness may be due to the stabilization and protection of BoNT by the NAPs.[15,16]

The most prominent of the NAPs is HA33, comprising up to ~30% of the 19S progenitor toxin complex mass, with binding specificity for diverse carbohydrates, depending on the specific *Clostridium* serotype and strain. For instance, the HA33 protein of serotype A (HA33/A) binds glycolipids and glycoproteins containing galactose.[13,17,18] In comparison, the HA components associated with BoNT/C contain two distinct carbohydrate-binding proteins, HA33/C (also referred to as type C HA1) and the C terminus of HA70/C (HA3b) that both recognize sialic acid-containing glycolipids and glycoproteins, albeit with different specificities.[19] Little is known about the function of the other NAPs, HA17 and NTNHA. Here, we report the crystal structure of HA33/A from *C. botulinum*, the first structure of a serotype A NAP.

Results and Discussion

HA33/A structure

HA33/A was isolated from the progenitor toxin complex of the *C. botulinum* Hall strain and its X-ray crystal structure (Figure 1) was determined to 1.50 Å resolution by molecular replacement using the HA33/C structure (PDB code 1QXM[17]) as the search model. Data collection, model building and refinement statistics are summarized in Table 1. The final model includes two protein molecules, chains A and B (residues 10–293), and 564 water molecules in the asymmetric unit. The final *R*-factor is 16.8% with an R_{free} factor of 19.4%. The Ramachandran plot, produced by MolProbity,[20] shows that all residues lie within allowed regions. No electron density was observed for the first nine N-terminal residues. This finding is consistent with the observed post-translational modification, which removes the first five N-terminal amino acid residues.[21] It is not known whether this modification has any functional or serotype-specific repercussions for HA33/A.

Each HA33/A molecule (Figure 1) consists of a single polypeptide chain of 284 (10–293) residues with an overall shape reminiscent of a dumbbell. HA33/A contains two β-trefoil domains connected by a short α-helix. The dimensions of the HA33/A monomer are 70 Å×40 Å×37 Å, with an overall surface area of 10,800 Å^2. The total β-strand and α-helical content is 48% and 10%, respectively, which is considerably lower as compared to the β-strand content predicted by FT-IR and far-UV circular dichroism (74–77%).[22] Each β-trefoil domain consists of three homologous β-trefoil repeats that are arranged about a pseudo 3-fold axis to form a 12-stranded anti-parallel β-barrel capped by three β-hairpins. Each β-trefoil repeat is composed of four β-strands with 1234 topology, with the second and third strand separated by a β-hairpin and the other two strands connected by loops of variable length. These repeats, designated 1α, 1β, and 1γ for the N-terminal domain and 2α, 2β, and 2γ for the C-terminal domain, are composed of residues 10–55, 56–102, 103–144, 151–197, 198–245, and 246–293, respectively. The two domains of HA33/A are highly similar to each other and can be superposed with a C^{α} RMSD of 1.07 Å for 137 structurally equivalent residues and 24% sequence identity, according to the program TOP.[23]

Inter-domain conformational plasticity

The structures of the two HA33/A molecules present in the crystallographic asymmetric unit,

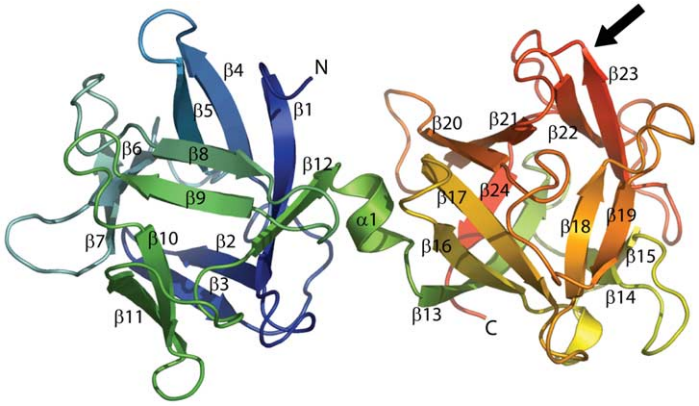

Figure 1. Crystal structure of HA33/A. Ribbon diagram of *Clostridium botulinum* HA33/A Hall-A strain color-coded from N terminus (blue) to C terminus (red) showing the two domains connected by a short helical linker. The β-strands (β1-β24) and α-helix (α1) are labeled. The putative carbohydrate-binding site is indicated with an arrow.

Table 1. Summary of crystallographic parameters, data collection and refinement statistics for HA33/A (PDB: 1YBI)

Data collection	
Space group	$P2_12_12$
Unit cell parameters (Å)	$a=104.08$, $b=146.60$, $c=35.71$
Wavelength (Å)	0.9179
Resolution range (Å)	25.00–1.50
Number of observations	158,552
Number of reflections	88,654
Completeness (%)	87.0 (78.6)[a]
Mean $I/\sigma(I)$	13.6 (3.7)[a]
R_{sym} on I	0.066 (0.299)[a]
Sigma cutoff	0.0
Highest resolution shell (Å)	1.53–1.50
Model and refinement statistics	
Resolution range (Å)	21.22–1.50
No. of reflections (total)	77,217
No. of reflections (test)	3820
Completeness (% total)	87.1
R_{cryst}/R_{free}	0.168/0.194
Stereochemical parameters	
Restraints (RMS observed)	
Bond lengths (Å)	0.009
Bond angles (deg.)	1.24
Average isotropic B-value (Å^2)	23.2
ESU based on R value (Å)	0.085
Protein residues/atoms	568/4677
Solvent molecules	564

ESU, estimated standard uncertainties;[47,51] $R_{sym} = \Sigma|I_i - \langle I_i\rangle|/\Sigma|I_i|$ where I_i is the scaled intensity of the ith measurement, and is the mean intensity for that reflection. $R_{cryst} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$ where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively. R_{free} as for R_{cryst}, but for 5.0% of the total reflections chosen at random and omitted from refinement.
[a] Highest resolution shell.

chains A and B, are slightly different, most likely due to domain flexibility, as indicated by a C^{α} RMSD of 1.91 Å. However, the individual N and C-terminal domains found in these two molecules are more similar and superpose with a C^{α} RMSD of 0.52 Å and 0.89 Å, respectively, with the largest differences in the C-terminal domains occurring at loops 223–228 and 243–248. The major disparity in the overall structures of the two molecules is due to the alternate packing of the two trefoil domains relative to the connecting helix linker that causes an approximate 10° rotation of the C-terminal domain with respect to the N-terminal domain (Figure 2(a)). The focal point of the domain rotation is located

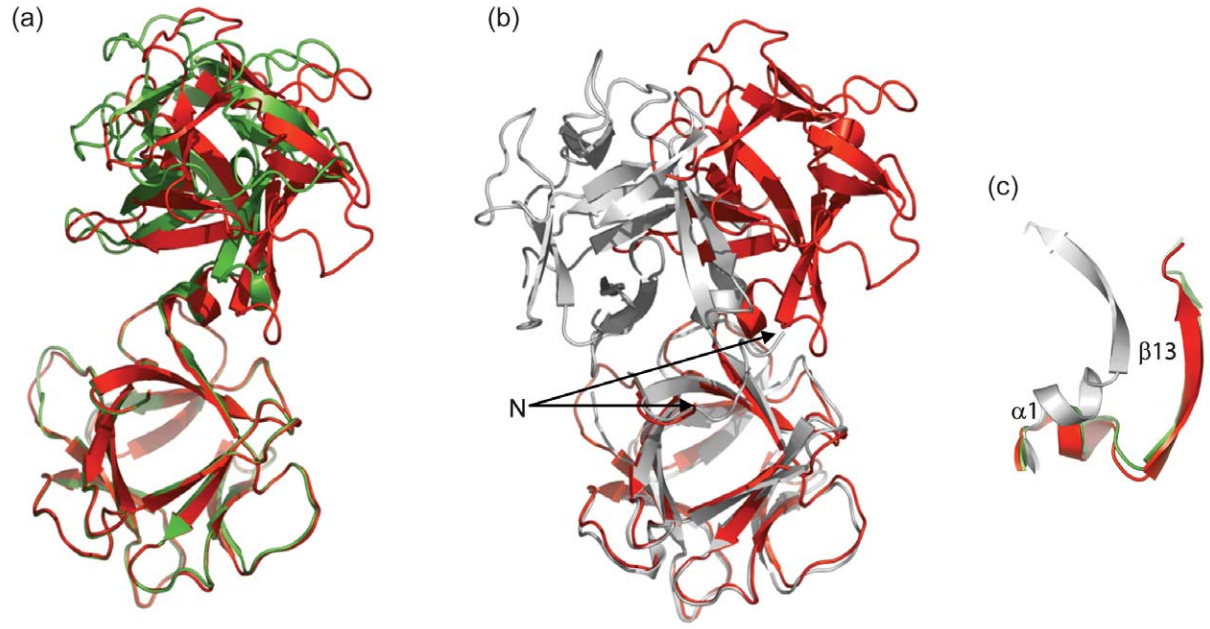

Figure 2. (a) Ribbon diagram of an N-terminal superposition of HA33/A chain A (red) and HA33/A chain B (green). (b) Same superposition as (a), but superposition of HA33/A chain A (red) and HA33/C (white). (c) Same superposition used above, but containing a close-up view of the helical linker region.

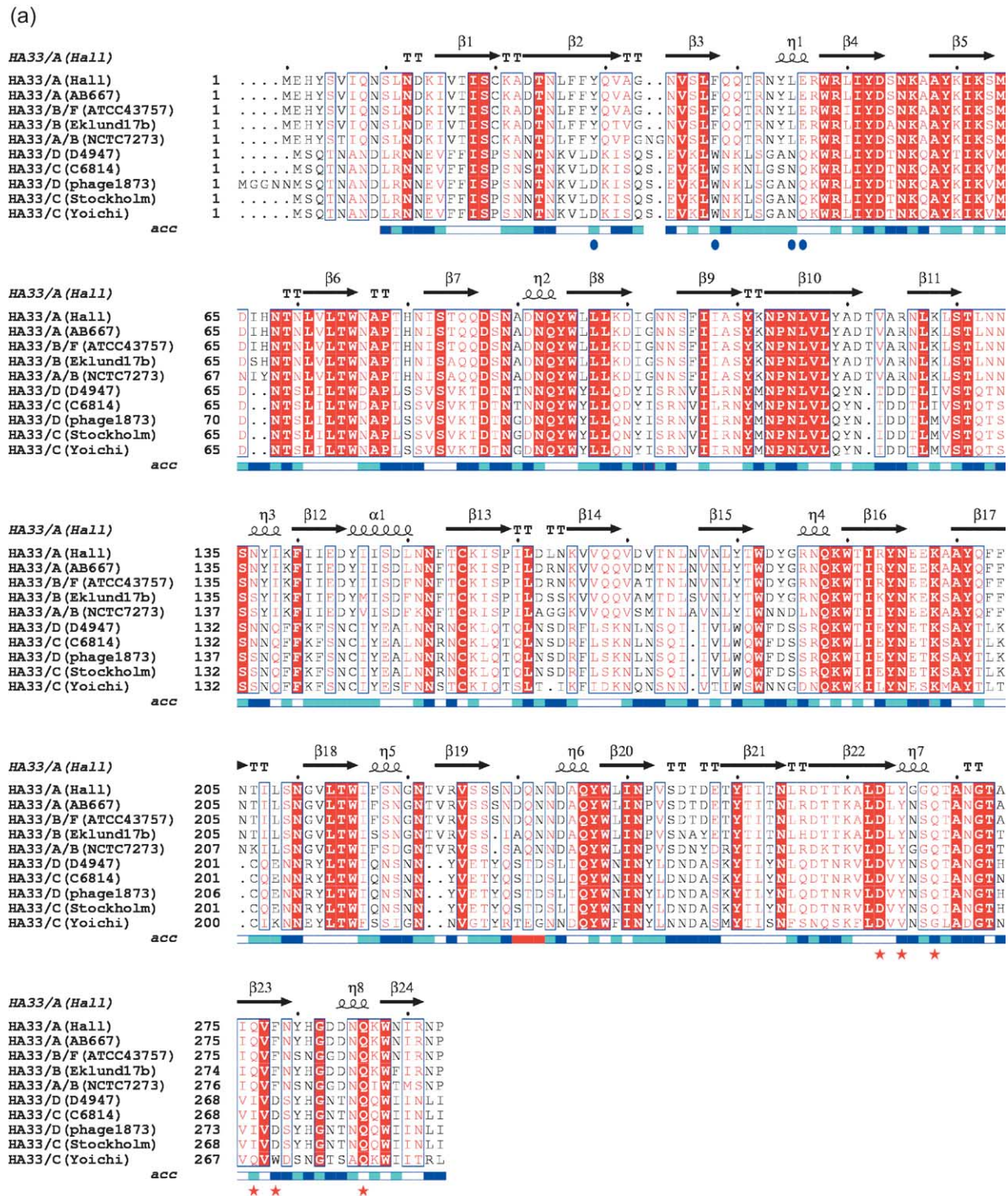

Figure 3(a) (*legend next page*)

after the hydrophobic linker helix α1 (residues 145–150) that balances between the two conformations observed in the two HA33/A chains. In the A chain, the inter-domain interactions are less extensive and mostly hydrophilic, such as the water-facilitated hydrogen-bonding of Glu143 and Glu196. While in the B chain hydrophobic interactions predominate, with Ile146 and Leu150 of the linker helix packing more deeply into a hydrophobic pocket of the C-terminal domain formed by residues Ile192, Ile240, Pro242 and Tyr250 than that observed in chain A. Additional interactions in the B chain include a bifurcated hydrogen-bonding network between the side-chain of Asp144 of the N-terminal domain with the amide nitrogen atoms of Ile146 and Ile147 of helix α1 and an inter-domain salt-bridge between Arg47 and Asp247, which are not found in the chain A molecule or the structure of HA33/C. Results obtained from a normal mode analysis using the

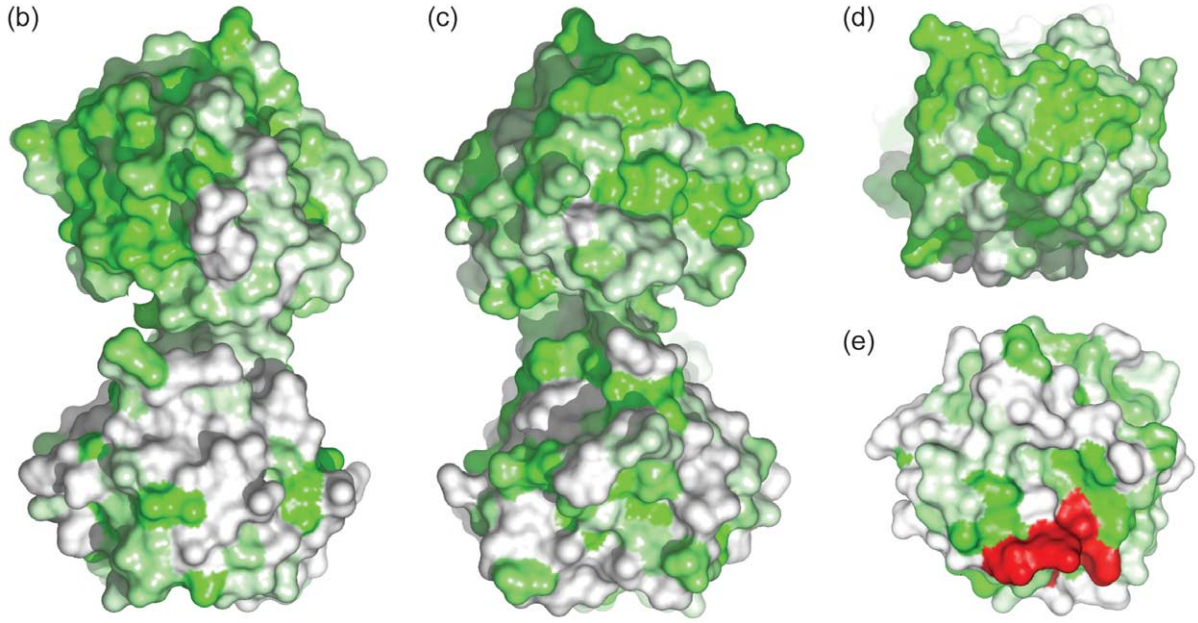

Figure 3. (a) Sequence alignment of HA33 homologs. The alignment shows strict sequence conservation in white letters and red background, and strong sequence conservation in red letters. The secondary structure elements of the HA33/A structure are labeled α (α-helix), η (3_{10} helix), β (β-strand) and TT (turn). The solvent-accessibility of each residue in the HA33/A structure is indicated in the bar at the base of the sequences, with white representing buried residues, dark blue representing solvent-accessible residues and light blue representing an intermediate value. The residues at the putative carbohydrate-binding sites of HA33/A and HA33/C are indicated underneath with red stars and blue circles, respectively. This Figure was prepared using ESPript.[49] (b) Surface representation of HA33/A showing conserved patches in green among HA33 homologs used in the sequence alignment in (a) indicating that the N-terminal domain (top) is more highly conserved than the C-terminal domain (bottom); (c) same as (b), but rotated 180°; (d) same as (b), but looking down the N-terminal domain; and (e) same as (b), but looking down the C-terminal domain with the residues Asp263, Tyr265, Gln276, Phe278, and Asn285 forming the putative carbohydrate recognition site in red. This Figure was prepared using the ConSurf server.[50]

ElNemo server[24] confer HA33/A flexibility between the trefoil domains at the helical linker. However, this inter-domain conformational flexibility is not seen in HA33/C, as the two molecules in its asymmetric unit superpose with a C^{α} RMSD of 0.53 Å. In this light, further studies are necessary to conclude whether the two conformations observed in the X-ray structure of HA33/A are functionally important or merely a crystal packing artifact.

Of particular note is that the inter-domain arrangement of the HA33/A molecules differs significantly from that observed for the HA33/C structure.[17] A comparison of the two serotype structures reveals an RMSD of 2.6 Å over 159 aligned residues (out of the 284 possible residues) with 36% sequence identity.[25] An approximately 60° rotation of the C-terminal domain was observed in the HA33/C structure as compared to the two HA33/A molecules, despite the fact that the N-terminal domains of these two serotypes superpose closely with an RMSD of 0.52 Å (Figure 2(b) and (c)). The dissimilarity of the structures found for these two serotypes is focused immediately before the linker helix α1, with the result that the two trefoil domains of HA33/C congregate together. This domain orientation dissimilarity may be serotype-dependent, since HA33/C has a longer N terminus located at the interface of the β-trefoil domains that does not undergo the post-translational cleavage (Figure 2(b)). The difference in the N termini may contribute to the serotype size differences in the progenitor toxin complexes, particularly since the 19 S progenitor toxin complex is produced only by serotype A and not by the other serotypes. N-terminal sequence analysis has revealed noteworthy serotype differences indicating that HA33/C[26] and HA33/D[27] do not undergo processing like that observed for HA33/A and HA33/B, which are similarly proteolytically shortened at their N termini.[21] Since BoNT serotypes A and B are both involved in human botulism, the high level of sequence conservation of their HA33 s may reflect their similar specificities, activities, and immuno-responses. Interestingly, the less toxic serotype A2,[3] E[28] and F[29] strains lack the genes encoding the HA components, and thus produce only 300 kDa 12 S toxin and have no ability to assemble with HA proteins while serotype G lacks only the HA33 gene.[30]

Sequence conservation of HA33 serotypes

An alignment of sequences containing HA33/A and nine more from other HA33 serotypes and strains is shown along with the secondary structure of HA33/A in Figure 3(a). There is substantially

greater sequence conservation observed at the surface of the N-terminal domain as compared to the C-terminal domain, though the majority of the conserved residues are solvent-inaccessible and are presumably responsible for maintaining the hydrophobic core of the β-trefoil fold (Figure 3(a)–(e)), with the greatest conservation being localized between the 1α and 1β repeats (β-strands β4–β6) possessing 65% identity. Since it is not clear which of the NAP components interact with the BoNT, the greater sequence conservation in the N-terminal domain of HA33/A suggests that this region is likely to be important for protein–protein interactions in the progenitor toxin complexes. Previous studies have reported that HA33/A accounts for most of the immunogenic response of the progenitor toxin complexes,[31] indicating that at least part of the molecule is exposed in the complex. Furthermore, C-terminally truncated variants of HA33/C lose their hemagglutination and erythrocyte-binding activity, suggesting that the C-terminal domain contains the sugar-binding site.[18] The lower level of sequence conservation in the C-terminal domain and the fact that this domain likely possesses the carbohydrate-binding site collectively suggest that the C-terminal HA33 domain is solvent-exposed and the likely major contributor to the distinct antigenicity of the serotypes.

Putative HA33/A carbohydrate-binding site

Recently, HA33/A has been proposed to contain a single sugar-binding site at the 2γ trefoil repeat specific for carbohydrates containing galactose, as determined by isothermal titration calorimetry and mutagenesis.[17] In many instances, the β-trefoil fold has been associated with oligosaccharide-binding ability, leading to a characteristic HA activity, including the plant toxin ricin, a prototypical ganglioside-binding protein, and other proteins like the ribosome-inactivating and potential cancer therapeutic mistletoe lectin I. The ricin and mistletoe lectin I structures revealed a domain architecture that are similar to HA33/A with two β-trefoil domains. The complex crystal structures of ricin bound to lactose and mistletoe lectin I bound to galactose revealed that only the 1α and 2γ repeats of these two proteins are involved in carbohydrate recognition.[32,33] Based on the structural comparison of HA33/A to ricin (PDB code 2AAI, with an overall sequence identity of 14%), the 2γ trefoil repeat (25% identity) of the HA33/A C-terminal domain likely contains the sugar-recognition site. An overlay of the residues of the 2γ repeat in the HA33/A structure with the residues in the homologous ricin lectin chain shows a spatial coincidence in the lactose-binding site when residues contained in a 4 Å sphere around lactose are overlaid (Figure 4(a)). The key residues of Asp234, Arg235, Glu239, Tyr248, and Asn255 of ricin have counterparts in HA33/A residues Asp263, Tyr265, Gln268, Phe278, and Asn285, respectively. This putative carbohydrate-binding site in HA33/A contains three structurally essential components for galactoside-binding that have been proposed for the ricin lectin chain.[33,34] In support of the 2γ sugar-binding site for HA33/A are mutants Asp263 and Asn285 to Ala that lose their ability to bind carbohydrates.[17]

A comparison of the 2γ repeat of HA33/C with the equivalent region in HA33/A (Figure 4(b)) shows that they are highly similar, suggesting that HA33/C contains the necessary components for carbohydrate binding, but with noted differences that are likely to be important for ligand discrimination of *N*-acetylneuraminic acid-containing moieties by HA33/C. The residues of Asp263, Tyr265, Gln268, and Asn285 in HA33/A responsible for the putative carbohydrate recognition have conserved counterparts in HA33/C residues Asp256, Tyr258, Gln261, and Asn278, respectively. In addition, within strand β23 at location 278 in HA33/A (Hall) (Figure 3(a)) which has a Phe

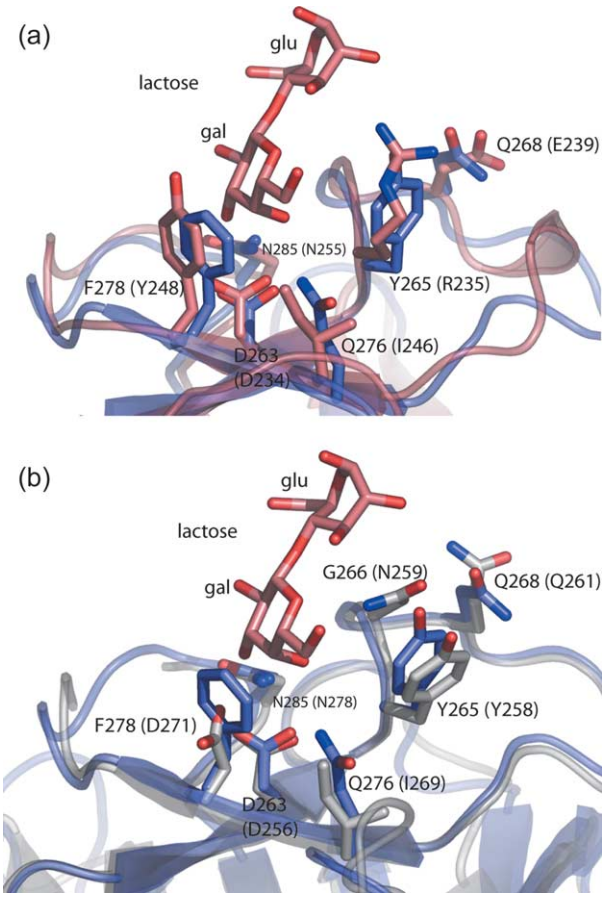

Figure 4. (a) Close-up view of HA33/A super-posed on the ricin lactose-binding site at the 2γ repeat region. Putative carbohydrate-binding residues as observed in HA33/A (slate blue) and their counterparts as found in the lactose-bound ricin structure (salmon). Residue labels are indicated for HA33/A with those from the ricin structure in parentheses. (b) Similar overlay as in (a), but a close-up view at the 2γ repeat of HA33/A (slate blue) superposed on HA33/C (white), with a modeled lactose molecule (salmon) and with HA33/C labels in parentheses.

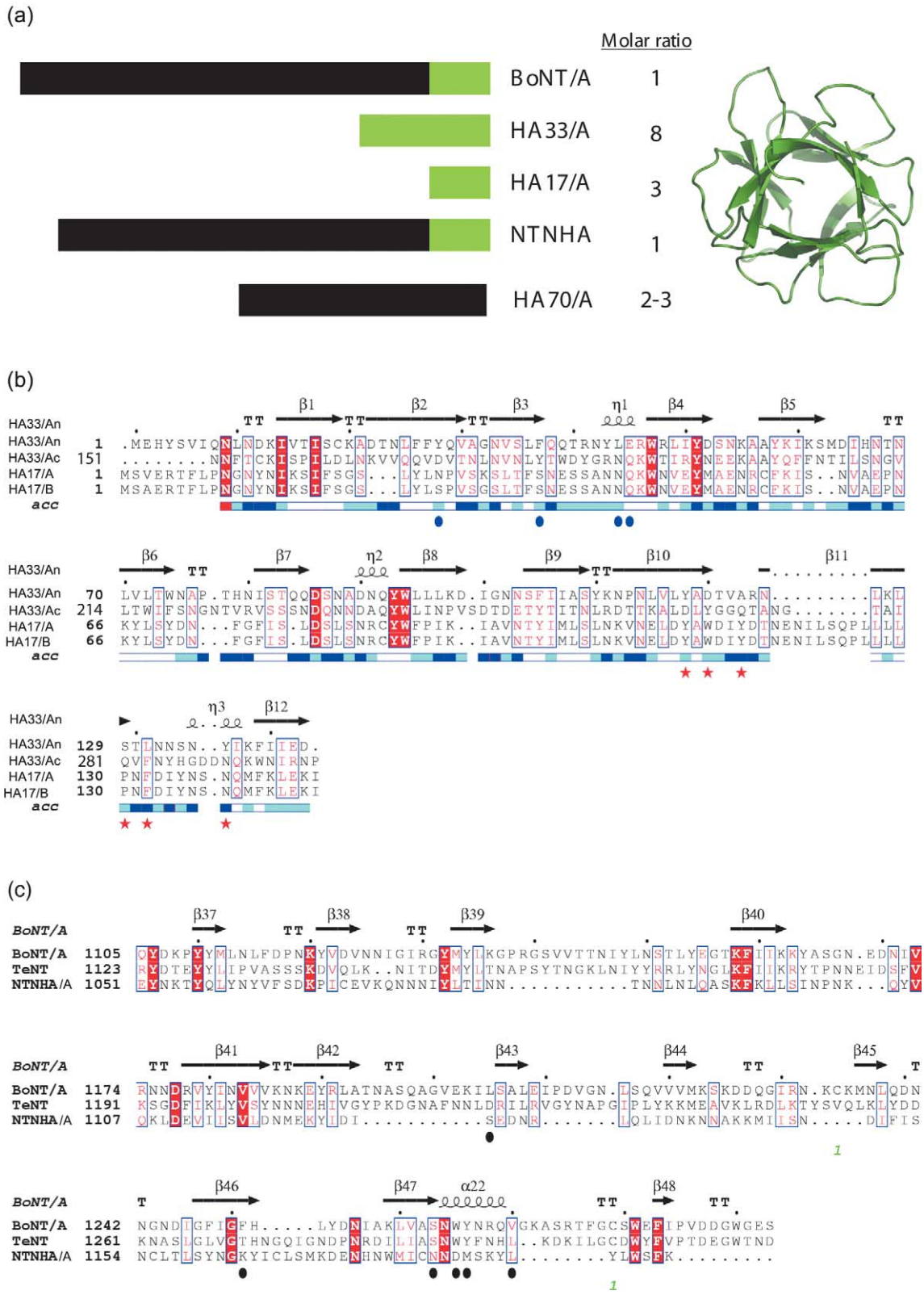

Figure 5. (a) Scheme showing the β-trefoil content in the components of the 900 kDa progenitor toxin complex based on the molecular composition of the type A toxin reported,[11] although the exact stoichiometry is still under debate. (b) Sequence alignment of HA17/A with HA33/A N and C-terminal trefoil domains. The alignment shows strict sequence conservation in white letters and red background, and strong sequence conservation in red letters. The secondary structure elements of the HA33/A N-terminal domain structure are labeled α (α-helix), η (3_{10} helix), β (β-strand) and TT (turn). The solvent accessibility of each residue is indicated in the bar displayed at the base of the sequences, with white representing buried residues, dark blue representing solvent-accessible residues and light blue representing an intermediate value. The residues at the putative carbohydrate-binding sites of HA33/A and HA33/C are indicated underneath with red stars and blue circles, respectively. (c) Same as (b), but alignment of NTNHA with the β-trefoil domains of TeNT and BoNT/A (Hall-A strain) with the secondary structure elements of the BoNT/A (PDB code 3BTA). The residues at the receptor-binding sites of TeNT are indicated underneath with black circles.

residue, the sequences from the type C and D HA33s (except Yoichi) have an Asp residue. This variation might account for the differences in the types of carbohydrates that bind to these two groups of HA33s (i.e. A, B *versus* C, D). Further support for the presence of a carbohydrate-binding site at the 2γ repeat in HA33/C includes the fact that a C-terminally truncated variant of HA33/C lacking the 2γ repeat loses sugar-binding activity.[18]

A superposition of the HA33 1α repeats for the A and C serotypes (not shown) reveals that HA33/A is not expected to bind carbohydrates at this site. Only the 2γ HA33/A β-trefoil motif possesses all of the structural elements required for carbohydrate recognition. The absence of a second carbohydrate-binding site indicates that the HA activity of the 16 S and 19 S progenitor toxin complexes requires HA33/A oligomerization, either with itself or with other NAPs, in order to create multivalent sugar-binding sites as previously suggested.[35]

Role of HA33/A protein in the progenitor toxin complexes

Previous studies suggest that the HA33/A protein is a dimer in solution as determined by gel filtration and mass spectrometry.[22] Secondly, analysis of the protein stoichiometry in the 16 S and 19 S progenitor toxins complexes led to the hypothesis that a HA33/A dimer links two molecules of 16 S neurotoxin complex to form the 19 S progenitor toxin complex.[11,36] However, an analysis of the crystal packing of the HA33/A molecules reveals only limited surface contacts between the two molecules. This interface accounts for only 10.1% of the buried surface area, consisting of mostly hydrophilic interactions such as residues Gln34, Asn76, Pro78, Thr79, Asn81, Gln85, His90, Lys128, and Thr131 of the N-terminal domain from one HA33/A molecule interacting with residues Met64, Ile66, His67, Asp245, and Asp247 of both domains from the other HA33/A molecule. Given such a small and weak dimer interface, it is unlikely that the crystallographic interface is physiologically significant and is unlikely to cross-link two molecules of 16 S toxin to form the 19 S progenitor toxin complex. Furthermore, to our knowledge the crystallographic dimer interface is unlike that seen in other β-trefoil-containing structures. Nonetheless, further functional studies are needed to conclusively determine the structural role of HA33/A in forming the assembled progenitor toxin complexes.

Reoccurring β-trefoil fold in the progenitor toxin complex

In addition to the β-trefoil domains found in HA33, BoNT/A and BoNT/B have previously been shown to contain a single β-trefoil fold at the C-terminal binding domain of the heavy chain.[37,38] Fold recognition methods also detect this β-trefoil domain in the NAPs HA17 and NTNHA (FFAS scores −23 and −208, respectively; scores below −9.5 typically indicate significant similarity with less than 3% of false positives).[39] Based on this analysis, the β-trefoil motif collectively forms nearly half of the mass of the 900 kDa progenitor toxin serotype A complex (Figure 5(a)). The importance of sequence conservation of BoNT with this fold and ganglioside recognition has been reported;[40] however, the significance of the trefoil fold within the context of the NAPs and the progenitor complex has yet to be addressed. Based on the HA33/A structure, a sequence alignment of HA17/A with the N and C-terminal domains of HA33/A (each with an overall sequence similarity of 28%) allows the β-trefoil repeats in HA17/A to be identified and assessed for sugar-binding conservation (Figure 5(b)). Most of the sequence conservation between HA17/A and HA33/A N and C-terminal domains is located in the γ repeats. Three residues that are important for carbohydrate recognition in the HA33/A C-terminal domain Asn263, Tyr265, and Asn285 are conserved with Asn113, Tyr115, and Asn138 in HA17/A. However, the three other residues of Gln268, Gln277, and Phe279 of HA33/A that form the rest of the carbohydrate-binding site have no equivalent counterpart in Ile121, Leu129, and Asn131 of HA17/A suggesting that HA17/A lacks the necessary molecular requirements for sugar-binding. Further support that HA17 does not bind carbohydrates comes from experiments with GST fusion proteins of HA17/A and HA17/C, which did not bind to erythrocytes and intestinal microvilli.[12,19] In addition, it has been reported that NTNHA also does not possess HA activity and that it does not bind to erythrocytes, but NTNHA is a critical component in formation of HA-positive progenitor toxin complexes.[43] NTNHA shows significant sequence similarity to the β-trefoil-containing binding domains of BoNT/A and tetanus neurotoxin (TeNT) from *Clostridium tetani* with sequence identities in the α and β repeats of 16% and 18%, respectively, and an overall sequence similarity of 31%. The sequence comparison of their β-trefoil subdomains (Figure 5(c)) offers clues into why NTNHA does not bind carbohydrates. The crystal structure of TeNT bound to a GT1b ganglioside receptor analog (Gal4-GalNAc3) at the γ repeat provides a structural prototype for the characterization of the ganglioside binding site,[41] and has proven crucial in identifying the ganglioside-binding site for BoNT/A and BoNT/B.[37,42] The TeNT key residues of Ser1287, Trp1289, and Tyr1290 are conserved in BoNT/A with residues Ser1264, Trp1266, and Tyr1267, but are poorly conserved in NTNHA with corresponding residues of Asn1180, Asp1182, and Met1183.

The identification of the β-trefoil domain in a majority of the components of the 900 kDa serotype A progenitor toxin suggests this fold is a result of structural domain duplication. It also allows one to postulate models in determining the molecular composition of the components of the progenitor

toxin complexes. With most of the molecular pieces of the proverbial progenitor toxin puzzle now at hand, our future work will focus on determining the intimate protein–protein contacts that make up the progenitor toxin complexes essential for BoNT protection, uptake, and activation.

Materials and Methods

Protein production and crystallization

HA33/A from the *C. botulinum* Hall strain was purified as described,[7] and kindly provided by the DasGupta/Johnson laboratories at the University of Wisconsin. Briefly, the ammonium sulfate-precipitated protein was dissolved by dialysis against a buffer containing 10 mM sodium acetate (pH 5.5), 200 mM $(NH_4)_2SO_4$ and concentrated to 12 mg/ml by ultrafiltration. The protein was crystallized by the hanging-drop, vapor-diffusion method using equal volumes of protein and reservoir solution. The crystallization reservoir solution contained 20% (w/v) PEG 4000, 15% (v/v) isopropanol, 200 mM Li_2SO_4, and 0.1 M Hepes at pH 7.5. Crystals were stabilized in freshly prepared reservoir solution containing 30% isopropanol for approximately 15 seconds prior to cryo-cooling in liquid nitrogen. The crystals were indexed in the orthorhombic space group $P2_12_12$ (Table 1).

Data collection

Diffraction data were collected at Stanford Synchrotron Radiation Laboratory (SSRL, Stanford, USA) on beamline 9-1 (Table 1). Data were integrated, reduced, and scaled using Denzo and Scalepack.[44] Data statistics are summarized in Table 1.

Structure solution and refinement

Four homology models of the individual HA33/A β-trefoil domains were constructed with program Whatif,[45] based on the FFAS[41] alignment with HA33/C (PDB code 1QXM, sequence identity 38%). Multiple molecular replacement searches were carried out in program MOLREP[46] on a 80 CPU Linux cluster expecting four copies of a single HA33/A domain in the asymmetric unit. Out of 160 MR trials only four obtained with a model based on the N-terminal domain of the HA33/C (chain B) structure had values of R_{free} below 0.50 after rigid body and restrained refinement in Refmac5.[47] Subsequent manual rebuilding and refinement was carried out in programs O[48] and Refmac5. Refinement statistics are summarized in Table 1. The final model includes two protein molecules and 564 water molecules. No electron density was observed for residues 1–9. Analysis of the stereochemical quality of the model was accomplished using the AutoDepInputTool†. Figures were prepared with PYMOL (DeLano Scientific).

Ligand docking

The probable binding site for the carbohydrate substrate was obtained by superimposing the β-trefoil domain of the lactose-bound ricin structure[34] with the C-terminal trefoil domain of HA33/A using the program TOP.[23]

Sequence alignments

NAP sequences were aligned using FFAS[41] and ClustalW.[48] A gap opening penalty of ten, gap extension penalty of 0.05, and gap separation distance of eight were used with the BLOSUM62 matrix. The alignment Figures were prepared using ESPript using DSSP secondary structure assignments.[49]

Protein Data Bank accession code

Atomic coordinates and experimental structure factors of HA33/A have been deposited with the PDB and are accessible under the code 1YBI.

† http://deposit.pdb.org/adit/

Acknowledgements

We thank Angela Walker and Dr Marianne Patch for assistance in manuscript preparation and Bibhuti DasGupta, Eric Johnson and Bill Tepp at the University of Wisconsin for HA33 protein samples. This work was supported by contract DAMD17-00-C-0040 from the Department of the Army and, in part, by National Institutes of Health Protein Structure Initiative Grant GM62411 from the National Institute of General Medical Sciences (http://www.nigms.nih.gov). Portions of this research were carried out at the Stanford Synchrotron Radiation Laboratory (SSRL), a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health (National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences).

References

1. Simpson, L. L. (2004). Identification of the major steps in botulinum toxin action. *Annu. Rev. Pharmacol. Toxicol.* **44**, 167–193.
2. Verastegui, C., Lalli, G., Bohnert, S., Meunier, F. A. & Schiavo, G. (2002). Clostridial neurotoxins. *J. Toxicol.-Toxin Rev.* **21**, 203–227.
3. Johnson, E. A. & Bradshaw, M. (2001). *Clostridium botulinum* and its neurotoxins: a metabolic and cellular perspective. *Toxicon*, **39**, 1703–1722.
4. Humeau, Y., Doussau, F., Grant, N. J. & Poulain, B. (2000). How botulinum and tetanus neurotoxins block neurotransmitter release. *Biochimie*, **82**, 427–446.
5. Sugii, S., Ohishi, I. & Sakaguchi, G. (1977). Correlation between oral toxicity and invitro stability of *Clostridium botulinum* type-A and type-B toxins of different molecular sizes. *Infect. Immun.* **16**, 910–914.

6. Sharma, S. K. & Singh, B. R. (1998). Hemagglutinin binding mediated protection of botulinum neurotoxin from proteolysis. *J. Natural Toxin*, **7**, 239–253.
7. Fu, F. N., Sharma, S. K. & Singh, B. R. (1998). A protease-resistant novel hemagglutinin purified from type A *Clostridium botulinum*. *J. Protein Chem.* **17**, 53–60.
8. Chen, F., Kuziemko, G. M. & Stevens, R. C. (1998). Biophysical characterization of the stability of the 150-kilodalton botulinum toxin, the nontoxic component, and the 900-kilodalton botulinum toxin complex species. *Infect. Immun.* **66**, 2420–2425.
9. Popoff, M. R. & Marvaud, J. C. (1999). Structural and genomic features of clostridial neurotoxins. In *The Comprehensive Sourcebook of Bacterial Protein Toxins* (Alouf, J. E. & Freer, J. H., eds), 2nd edit., pp. 174–201, Academic Press, San Diego.
10. Popoff, M. R. (1998). Interactions between bacterial toxins and intestinal cells. *Toxicon*, **36**, 665–685.
11. Inoue, K., Fujinaga, Y., Watanabe, T., Ohyama, T., Takeshi, K., Moriishi, K. *et al*. (1996). Molecular composition of *Clostridium botulinum* type A progenitor toxins. *Infect. Immun.* **64**, 1589–1594.
12. Fujinaga, Y., Inoue, K., Nomura, T., Sasaki, J., Marvaud, J. C., Popoff, M. R., Kozaki, S. *et al*. (2000). Identification and characterization of functional subunits of *Clostridium botulinum* type A progenitor toxin involved in binding to intestinal microvilli and erythrocytes. *FEBS Letters*, **467**, 179–183.
13. Inoue, K., Fujinaga, Y., Honke, K., Arimitsu, H., Mahmut, N., Sakaguchi, Y. *et al*. (2001). *Clostridium botulinum* type A haemagglutinin-positive progenitor toxin (HA(+)-PTX) binds to oligosaccharides containing Gal beta 1-4GlcNAc through one subcomponent of haemagglutinin (HA1). *Microbiology*, **147**, 811–819.
14. Fujinaga, Y., Inoue, K., Watanabe, S., Yokota, K., Hirai, Y., Nagamachi, E. & Oguma, K. (1997). The haemagglutinin of *Clostridium botulinum* type C progenitor toxin plays an essential role in binding of toxin to the epithelial cells of guinea pig small intestine, leading to the efficient absorption of the toxin. *Microbiology*, **143**, 3841–3847.
15. Schantz, E. J. & Johnson, E. A. (1992). Properties and use of botulinum toxin and other microbial neurotoxins in medicine. *Microbiol. Rev.* **56**, 80–99.
16. Johnson, E. A. (1999). Clostridial toxins as therapeutic agents: benefits of nature's most toxic proteins. *Annu. Rev. Microbiol.* **53**, 551–575.
17. Inoue, K., Sobhany, M., Transue, T. R., Oguma, K., Pedersen, L. C. & Negishi, M. (2003). Structural analysis by X-ray crystallography and calorimetry of a haemagglutinin component (HA1) of the progenitor toxin from *Clostridium botulinum*. *Microbiology*, **149**, 3361–3370.
18. Sagane, Y., Kouguchi, H., Watanabe, T., Sunagawa, H., Inoue, K., Fujinaga, Y., Oguma, K. *et al*. (2001). Role of C-terminal region of HA-33 component of botulinum toxin in hemagglutination. *Biochem. Biophys. Res. Commun.* **288**, 650–657.
19. Fujinaga, Y., Inoue, K., Watarai, S., Sakaguchi, Y., Arimitsu, H., Lee, J. C., Jin, Y. L. *et al*. (2004). Molecular characterization of binding subcomponents of *Clostridium botulinum* type C progenitor toxin for intestinal epithelial cells and erythrocytes. *Microbiology*, **150**, 1529–1538.
20. Davis, I. W., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2004). MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucl. Acids Res.* **32**, W615–W619.
21. East, A. K., Stacey, J. M. & Collins, M. D. (1994). Cloning and sequencing of a hemagglutinin component of the botulinum neurotoxin complex encoded by *Clostridium botulinum* type-A and type-B. *System. Appl. Microbiol.* **17**, 306–312.
22. Sharma, S. K., Fu, F. N. & Singh, B. R. (1999). Molecular properties of a hemagglutinin purified from type A *Clostridium botulinum*. *J. Protein Chem.* **18**, 29–38.
23. Lu, G. G. (2000). TOP: a new method for protein structure comparisons and similarity searches. *J. Appl. Crystallog.* **33**, 176–183.
24. Suhre, K. & Sanejouand, Y. H. (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucl. Acids Res.* **32**, W610–W614.
25. Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**, 316–319.
26. Inoue, K., Fujinaga, Y., Honke, K., Yokota, K., Ikeda, T., Ohyama, T., Takeshi, K. *et al*. (1999). Characterization of haemagglutinin activity of *Clostridium botulinum* type C and D 16 S toxins, and one subcomponent of haemagglutinin (HA1). *Microbiology*, **145**, 2533–2542.
27. Ohyama, T., Watanabe, T., Fujinaga, Y., Inoue, K., Sunagawa, H., Fujii, N. & Oguma, K. (1995). Characterization of nontoxic–nonhemagglutinin component of the 2 types of progenitor toxin (M and L) produced by *Clostridium botulinum* type-D Cb-16. *Microbiol. Immun.* **39**, 457–465.
28. Fujii, N., Kimura, K., Yokosawa, N., Yashiki, T., Tsuzuki, K. & Ouma, K. (1993). The complete nucleotide sequence of the gene encoding the nontoxic component of *Clostridium botulinum* type E progenitor toxin. *J. Gen. Microbiol.* **139**, 79–86.
29. East, A. K. & Collins, M. D. (1994). Conserved structure of genes encoding components of botulinum neurotoxin complex-M and the sequence of the gene coding for the nontoxic component in nonproteolytic *Clostridium botulinum* type-F. *Curr. Microbiol.* **29**, 69–77.
30. Kouguchi, H., Watanabe, T., Sagane, Y., Sunagawa, H. & Ohyama, T. (2002). *In vitro* reconstitution of the *Clostridium botulinum* type D progenitor toxin. *J. Biol. Chem.* **277**, 2650–2656.
31. Sharma, S. K. & Singh, B. R. (2000). Immunological properties of HN-33 purified from type A *Clostridium botulinum*. *J. Natural Toxin*, **9**, 357–362.
32. Niwa, H., Tonevitsky, A. G., Agapov, I. I., Saward, S., Pfuller, U. & Palmer, R. A. (2003). Crystal structure at 3 Å of mistletoe lectin I, a dimeric type-II ribosome-inactivating protein, complexed with galactose. *Eur. J. Biochem.* **270**, 2739–2749.
33. Rutenber, E. & Robertus, J. D. (1991). Structure of ricin B-chain at 2.5-Å resolution. *Proteins: Struct. Funct. Genet.* **10**, 260–269.
34. Rutenber, E., Ready, M. & Robertus, J. D. (1987). Structure and evolution of ricin B-chain. *Nature*, **326**, 624–626.
35. Hazes, B. (1996). The (QxW)(3) domain: a flexible lectin scaffold. *Protein Sci.* **5**, 1490–1501.
36. Sakaguchi, G., Kozaki, S. & Ohishi, I. (1984). Structure and function of botulinum toxins. In *Bacterial Protein Toxins* (Alouf, J. E., Fahrenbach, F. J., Freer, J. H. & Jeljaszewicz, J., eds), pp. 435–443, Academic Press, London.

37. Swaminathan, S. & Eswaramoorthy, S. (2000). Structural analysis of the catalytic and binding sites of *Clostridium botulinum* neurotoxin B. *Nature Struct. Biol.* **7**, 693–699.
38. Lacy, D. B., Tepp, W., Cohen, A. C., DasGupta, B. R. & Stevens, R. C. (1998). Crystal structure of botulinum neurotoxin type A and implications for toxicity. *Nature Struct. Biol.* **5**, 898–902.
39. Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). Improving the quality of twilight-zone alignments. *Protein Sci.* **9**, 1487–1496.
40. Ginalski, K., Venclovas, C., Lesyng, B. & Fidelis, K. (2000). Structure-based sequence alignment for the beta-trefoil subdomain of the clostridial neurotoxin family provides residue level information about the putative ganglioside binding site. *FEBS Letters*, **482**, 119–124.
41. Fotinou, C., Emsley, P., Black, I., Ando, H., Ishida, H., Kiso, M. *et al*. (2001). The crystal structure of tetanus toxin Hc fragment complexed with a synthetic GT1b analogue suggests cross-linking between ganglioside receptors and the toxin. *J. Biol. Chem.* **276**, 32274–32281.
42. Rummel, A., Mahrhold, S., Bigalke, H. & Binz, T. (2004). The H-CC-domain of botulinum neurotoxins A and B exhibits a singular ganglioside binding site displaying serotype specific carbohydrate interaction. *Mol. Microbiol.* **51**, 631–643.
43. Fujita, R., Fujinaga, Y., Inoue, K., Nakajima, H., Kumon, H. & Oguma, K. (1995). Molecular characterization of 2 forms of nontoxic–nonhemagglutinin components of *Clostridium botulinum* type-A progenitor toxins. *FEBS Letters*, **376**, 41–44.
44. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Acta Crystallog. sect. A*, **276**, 307–326.
45. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56.
46. Vagin, A. & Teplyakov, A. (1997). MOLREP: an automated program for molecular replacement. *J. Appl. Crystallog.* **30**, 1022–1025.
47. Collaborative Computational Project Number 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallog. sect. D*, **50**, 760–763.
48. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Clustal-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting. Position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
49. Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.
50. Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor, D., Martz, E. & Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
51. Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of gammaB- and betaB2-crystallin. *Acta Crystallog. sect. D*, **54**, 243–252.

Edited by M. Guss

(Received 28 October 2004; received in revised form 15 December 2004; accepted 16 December 2004)